(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 August 2001 (16.08.2001)

PCT

(10) International Publication Number
WO 01/59066 A2

(51) International Patent Classification[7]: C12N

(21) International Application Number: PCT/US01/40091

(22) International Filing Date: 12 February 2001 (12.02.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/181,630	10 February 2000 (10.02.2000)	US
60/186,904	3 March 2000 (03.03.2000)	US
60/197,851	14 April 2000 (14.04.2000)	US

(71) Applicant *(for all designated States except US)*: **XENCOR, INC.** [US/US]; 111 W. Lemon Avenue, Monrovia, CA 91016 (US).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **DAHIYAT, Bassil, I.** [US/US]; 4413 San Andreas Avenue, Los Angeles, CA 90065 (US).

(74) Agent: **TRECARTIN, Richard, F.**; Flehr Hohbach Test Albritton & Herbert LLP, Suite 3400, 4 Embarcadero Center, San Francisco, CA 94111-4187 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/59066 A2

(54) Title: PROTEIN DESIGN AUTOMATION FOR PROTEIN LIBRARIES

(57) Abstract: The invention relates to the use of protein design automation (PDA) to generate computationally prescreened secondary libraries of proteins, and to methods and compositions utilizing the libraries.

PROTEIN DESIGN AUTOMATION FOR PROTEIN LIBRARIES

FIELD OF THE INVENTION

The invention relates to the use of a variety of computation methods, including protein design automation (PDA), to generate computationally prescreened secondary libraries of proteins, and to methods of making and methods and compositions utilizing the libraries.

BACKGROUND OF THE INVENTION

Directed molecular evolution can be used to create proteins and enzymes with novel functions and properties. Starting with a known natural protein, several rounds of mutagenesis, functional screening, and propagation of successful sequences are performed. The advantage of this process is that it can be used to rapidly evolve any protein without knowledge of its structure. Several different mutagenesis strategies exist, including point mutagenesis by error-prone PCR, cassette mutagenesis, and DNA shuffling. These techniques have had many successes; however, they are all handicapped by their inability to produce more than a tiny fraction of the potential changes. For example, there are 20^{500} possible amino acid changes for an average protein approximately 500 amino acids long. Clearly, the mutagenesis and functional screening of so many mutants is impossible; directed evolution provides a very sparse sampling of the possible sequences and hence examines only a small portion of possible improved proteins, typically point mutants or recombinations of existing sequences. By sampling randomly from the vast number of possible sequences, directed evolution is unbiased and broadly applicable, but inherently inefficient because it ignores all structural and biophysical knowledge of proteins.

In contrast, computational methods can be used to screen enormous sequence libraries (up to 10^{80} in a single calculation) overcoming the key limitation of experimental library screening methods such as directed molecular evolution. There are a wide variety of methods known for generating and evaluating sequences. These include, but are not limited to, sequence profiling (Bowie and Eisenberg, Science 253(5016): 164-70, (1991)), rotamer library selections (Dahiyat and Mayo, Protein Sci 5(5): 895-903 (1996); Dahiyat and Mayo, Science 278(5335): 82-7 (1997); Desjarlais and Handel, Protein

Science 4: 2006-2018 (1995); Harbury et al, PNAS USA 92(18): 8408-8412 (1995); Kono et al., Proteins: Structure, Function and Genetics 19: 244-255 (1994); Hellinga and Richards, PNAS USA 91: 5803-5807 (1994)); and residue pair potentials (Jones, Protein Science 3: 567-574, (1994)).

In particular, U.S.S.N.s 60/061,097, 60/043,464, 60/054,678, 09/127,926 and PCT US98/07254 describe a method termed "Protein Design Automation", or PDA, that utilizes a number of scoring functions to evaluate sequence stability.

It is an object of the present invention to provide computational methods for prescreening sequence libraries to generate and select secondary libraries, which can then be made and evaluated experimentally.

SUMMARY OF THE INVENTION

In accordance with the objects outlined above, the present invention provides methods for generating a secondary library of scaffold protein variants comprising providing a primary library comprising a rank-ordered list of scaffold protein primary variant sequences. A list of primary variant positions in the primary library is then generated, and a plurality of the primary variant positions is then combined to generate a secondary library of secondary sequences.

In an additional aspect, the invention provides methods for generating a secondary library of scaffold protein variants comprising providing a primary library comprising a rank-ordered list of scaffold protein primary variant sequences, and generating a probability distribution of amino acid residues in a plurality of variant positions. The plurality of the amino acid residues is combined to generate a secondary library of secondary sequences. These sequences may then be optionally synthesized and tested, in a variety of ways, including multiplexing PCR with pooled oligonucleotides, error prone PCR, gene shuffling, etc.

In a further aspect, the invention provides compositions comprising a plurality of secondary variant proteins or nucleic acids encoding the proteins, wherein the plurality comprises all or a subset of the secondary library. The invention further provides cells comprising the library, particularly mammalian cells.

In an additional aspect, the invention provides methods for generating a secondary library of scaffold protein variants comprising providing a first library rank-ordered list of scaffold protein primary variants;

generating a probability distribution of amino acid residues in a plurality of variant positions; and synthesizing a plurality of scaffold protein secondary variants comprising a plurality of the amino acid residues to form a secondary library. At least one of the secondary variants is different from the primary variants.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 depicts the synthesis of a full-length gene and all possible mutations by PCR. Overlapping oligonucleotides corresponding to the full-length gene (black bar, Step 1) are synthesized, heated and annealed. Addition of *Pfu* DNA polymerase to the annealed oligonucleotides results in the 5' →3' synthesis of DNA (Step 2) to produce longer DNA fragments (Step 3). Repeated cycles of heating, annealing (Step 4) results in the production of longer DNA, including some full-length molecules. These can be selected by a second round of PCR using primers (arrowed) corresponding to the end of the full-length gene (Step 5).

Figure 2 depicts the reduction of the dimensionality of sequence space by PDA screening. From left to right, 1: without PDA; 2: without PDA not counting Cysteine, Proline, Glycine; 3: with PDA using the 1% criterion, modeling free enzyme; 4: with PDA using the 1% criterion, modeling enzyme-substrate complex; 5: with PDA using the 5% criterion modeling free enzyme; 6: with PDA using the 5% criterion modeling enzyme-substrate complex.

Figure 3 depicts the active site of *B. circulans* xylanase. Those positions included in the PDA design are shown by their side chain representation. In red are wild type residues (their conformation was allowed to change, but not their amino acid identity). In green are positions whose conformation and identity were allowed to change (to any amino acid except proline, cysteine and glycine).

Figure 4 depicts cefotaxime resistance of *E. coli* expressing wild type (WT) and PDA Screened β-lactamase; results shown for increasing concentrations of cefotaxime.

Figure 5 depicts a preferred scheme for synthesizing a library of the invention. The wild-type gene, or any starting gene, such as the gene for the global minima gene, can be used. Oligonucleotides

comprising different amino acids at the different variant positions can be used during PCR using standard primers. This generally requires fewer oligonucleotides and can result in fewer errors.

Figure 6 depicts and overlapping extension method. At the top of Figure 6 is the template DNA showing the locations of the regions to be mutated (black boxes) and the binding sites of the relevant primers (arrows). The primers R1 and R2 represent a pool of primers, each containing a different mutation; as described herein, this may be done using different ratios of primers if desired. The variant position is flanked by regions of homology sufficient to get hybridization. In this example, three separate PCR reactions are done for step 1. The first reaction contains the template plus oligos F1 and R1. The second reaction contains template plus F2 and R2, and the third contains the template and F3 and R3. The reaction products are shown. In Step 2, the products from Step 1 tube 1 and Step 1 tube 2 are taken. After purification away from the primers, these are added to a fresh PCR reaction together with F1 and R4. During the Denaturation phase of the PCR, the overlapping regions anneal and the second strand is synthesized. The product is then amplified by the outside primers. In Step 3, the purified product from Step 2 is used in a third PCR reaction, together with the product of Step 1, tube 3 and the primers F1 and R3. The final product corresponds to the full length gene and contains the required mutations.

Figure 7 depicts a ligation of PCR reaction products to synthesize the libraries of the invention. In this technique, the primers also contain an endonuclease restriction site (RE), either blunt, 5' overhanging or 3' overhanging. We set up three separate PCR reactions for Step 1. The first reaction contains the template plus oligos F1 and R1. The second reaction contains template plus F2 and R2, and the third contains the template and F3 and R3. The reaction products are shown. In Step 2, the products of step 1 are purified and then digested with the appropriate restriction endonuclease. The digestion products from Step 2, tube 1 and Step 2, tube 2 and ligate them together with DNA ligase (step 3). The products are then amplified in Step 4 using primer F1 and R4. The whole process is then repeated by digesting the amplified products, ligating them to the digested products of Step 2, tube 3, and then amplifying the final product by primers F1 and R3. It would also be possible to ligate all three PCR products from Step 1 together in one reaction, providing the two restriction sites (RE1 and RE2) were different.

Figure 8 depicts blunt end ligation of PCR products. In this technique, the primers such as F1 and R1 do not overlap, but they abut. Again three separate PCR reactions are performed. The products from tube 1 and tube 2 are ligated, and then amplified with outside primers F1 and R4. This product is then ligated with the product from Step 1, tube 3. The final products are then amplified with primers F1 and R3.

Figure 9 depicts M13 single stranded template production of mutated PCR products. Primer1 and Primer2 (each representing a pool of primers corresponding to desired mutations) are mixed with the M13 template containing the wildtype gene or any starting gene. PCR produces the desired product (11) containing the combinations of the desired mutations incorporated in Primer1 and Primer2. This scheme can be used to produce a gene with mutations, or fragments of a gene with mutations that are then linked together via ligation or PCR for example.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is directed to methods of using computational screening of protein sequence libraries (that can comprise up to 10^{80} or more members) to select smaller libraries of protein sequences (that can comprise up to 10^{13} members), that can then be used in a number of ways. For example, the proteins can be actually synthesized and experimentally tested in the desired assay, for improved function and properties. Similarly, the library can be additionally computationally manipulated to create a new library which then itself can be experimentally tested.

The invention has two broad uses; first, the invention can be used to prescreen libraries based on known scaffold proteins. That is, computational screening for stability (or other properties) may be done on either the entire protein or some subset of residues, as desired and described below. By using computational methods to generate a threshold or cutoff to eliminate disfavored sequences, the percentage of useful variants in a given variant set size can increase, and the required experimental outlay is decreased.

In addition, the present invention finds use in the screening of random peptide libraries. As is known, signaling pathways in cells often begin with an effector stimulus that leads to a phenotypically describable change in cellular physiology. Despite the key role intracellular signaling pathways play in disease pathogenesis, in most cases, little is understood about a signaling pathway other than the initial stimulus and the ultimate cellular response.

Historically, signal transduction has been analyzed by biochemistry or genetics. The biochemical approach dissects a pathway in a "stepping-stone" fashion: find a molecule that acts at, or is involved in, one end of the pathway, isolate assayable quantities and then try to determine the next molecule in the pathway, either upstream or downstream of the isolated one. The genetic approach is classically a "shot in the dark": induce or derive mutants in a signaling pathway and map the locus by genetic crosses or complement the mutation with a cDNA library. Limitations of biochemical approaches

include a reliance on a significant amount of pre-existing knowledge about the constituents under study and the need to carry such studies out in vitro, post-mortem. Limitations of purely genetic approaches include the need to first derive and then characterize the pathway before proceeding with identifying and cloning the gene.

Screening molecular libraries of chemical compounds for drugs that regulate signal systems has led to important discoveries of great clinical significance. Cyclosporin A (CsA) and FK506, for examples, were selected in standard pharmaceutical screens for inhibition of T-cell activation. It is noteworthy that while these two drugs bind completely different cellular proteins -- cyclophilin and FK5O6 binding protein (FKBP), respectively, the effect of either drug is virtually the same -- profound and specific suppression of T-cell activation, phenotypically observable in T cells as inhibition of mRNA production dependent on transcription factors such as NF-AT and NF-κB. Libraries of small peptides have also been successfully screened in vitro in assays for bioactivity. The literature is replete with examples of small peptides capable of modulating a wide variety of signaling pathways. For example, a peptide derived from the HIV-1 envelope protein has been shown to block the action of cellular calmodulin.

Accordingly, generation of random or semi-random sequence libraries of proteins and peptides allows for the selection of proteins (including peptides, oligopeptides and polypeptides) with useful properties. The sequences in these experimental libraries can be randomized at specific sites only, or throughout the sequence. The number of sequences that can be searched in these libraries grows expontentially with the number of positions that are randomized. Generally, only up to 10^{12} - 10^{15} sequences can be contained in a library because of the physical constraints of laboratories (the size of the instruments, the cost of producing large numbers of biopolymers, etc.). Other practical considerations can often limit the size of the libraries to 10^{6} or fewer. These limits are reached for only 10 amino acid positions. Therefore, only a sparse sampling of sequences is possible in the search for improved proteins or peptides in experimental sequence libraries, lowering the chance of success and almost certainly missing desirable candidates. Because of the randomness of the changes in these sequences, most of the candidates in the library are not suitable, resulting in a waste of most of the effort in producing the library.

However, using the automated protein design techniques outlined below, virtual libraries of protein sequences can be generated that are vastly larger than experimental libraries. Up to 10^{80} candidate sequences can be screened computationally and those that meet design criteria which favor stable and functional proteins can be readily selected. An experimental library consisting of the favorable candidates found in the virtual library screening can then be generated, resulting in a much more efficient use of the experimental library and overcoming the limitations of random protein libraries.

Two principle benefits come from the virtual library screening: (1) the automated protein design generates a list of sequence candidates that are favored to meet design criteria; it also shows which positions in the sequence are readily changed and which positions are unlikely to change without disrupting protein stability and function. An experimental random library can be generated that is only randomized at the readily changeable, non-disruptive sequence positions. (2) The diversity of amino acids at these positions can be limited to those that the automated design shows are compatible with these positions. Thus, by limiting the number of randomized positions and the number of possibilities at these positions, the number of wasted sequences produced in the experimental library is reduced, thereby increasing the probability of success in finding sequences with useful properties.

In addition, by computationally screening very large libraries of mutants, greater diversity of protein sequences can be screened (i.e. a larger sampling of sequence space), leading to greater improvements in protein function. Further, fewer mutants need to be tested experimentally to screen a given library size, reducing the cost and difficulty of protein engineering. By using computational methods to pre-screen a protein library, the computational features of speed and efficiency are combined with the ability of experimental library screening to create new activities in proteins for which appropriate computational models and structure-function relationships are unclear.

Similarly, novel methods to create secondary libraries derived from very large computational mutant libraries allow the rapid testing of large numbers of computationally designed sequences.

In addition, as is more fully outlined below, the libraries may be biased in any number of ways, allowing the generation of secondary libraries that vary in their focus; for example, domains, subsets of residues, active or binding sites, surface residues, etc., may all be varied or kept constant as desired.

In general, as more fully outlined below, the invention can take on a wide variety of configurations. In general, primary libraries, e.g. libraries of all or a subset of possible proteins are generated computationally. This can be done in a wide variety of ways, including sequence alignments of related proteins, structural alignments, structural prediction models, databases, or (preferably) protein design automation computational analysis. Similarly, primary libraries can be generated via sequence screening using a set of scaffold structures that are created by perturbing the starting structure (using any number of techniques such as molecular dynamics, Monte Carlo analysis) to make changes to the protein (including backbone and sidechain torsion angle changes). Optimal sequences can be selected for each starting structures (or, some set of the top sequences) to make primary libraries.

Some of these techniques result in the list of sequences in the primary library being"scored", or "ranked" on the basis of some particular criteria. In some embodiments, lists of sequences that are generated without ranking can then be ranked using techniques as outlined below.

In a preferred embodiment, some subset of the primary library is then experimentally generated to form a secondary library. Alternatively, some or all of the primary library members are recombined to form a secondary library, e.g. with new members. Again, this may be done either computationally or experimentally or both.

Alternatively, once the primary library is generated, it can be manipulated in a variety of ways. In one embodiment, a different type of computational analysis can be done; for example, a new type of ranking may be done. Alternatively, and the primary library can be recombined, e.g. residues at different positions mixed to form a new, secondary library. Again, this can be done either computationally or experimentally, or both.

Accordingly, the present invention provides methods for generating secondary libraries of scaffold protein variants. By "protein" herein is meant at least two amino acids linked together by a peptide bond. As used herein, protein includes proteins, oligopeptides and peptides. The peptidyl group may comprise naturally occurring amino acids and peptide bonds, or synthetic peptidomimetic structures, i.e. "analogs", such as peptoids (see Simon *et al.*, PNAS USA 89(20):9367 (1992)).. The amino acids may either be naturally occurrng or non-naturally occuring; as will be appreciated by those in the art, any structure for which a set of rotamers is known or can be generated can be used as an amino acid. The side chains may be in either the (R) or the (S) configuration. In a preferred embodiment, the amino acids are in the (S) or L-configuration.

The scaffold protein may be any protein for which a three dimensional structure is known or can be generated; that is, for which there are three dimensional coordinates for each atom of the protein. Generally this can be determined using X-ray crystallographic techniques, NMR techniques, de novo modelling, homology modelling, etc. In general, if X-ray structures are used, structures at 2Å resolution or better are preferred, but not required.

The scaffold proteins may be from any organism, including prokaryotes and eukaryotes, with enzymes from bacteria, fungi, extremeophiles such as the archebacteria, insects, fish, animals (particularly mammals and particularly human) and birds all possible.

Thus, by "scaffold protein" herein is meant a protein for which a secondary library of variants is desired. As will be appreciated by those in the art, any number of scaffold proteins find use in the present invention. Specifically included within the definition of "protein" are fragments and domains of known proteins, including functional domains such as enzymatic domains, binding domains, etc., and smaller fragments, such as turns, loops, etc. That is, portions of proteins may be used as well. In addition, "protein" as used herein includes proteins, oligopeptides and peptides. In addition, protein variants, i.e. non-naturally occuring protein analog structures, may be used.

Suitable proteins include, but are not limited to, industrial and pharmaceutical proteins, including ligands, cell surface receptors, antigens, antibodies, cytokines, hormones, transcription factors, signaling modules, cytoskeletal proteins and enzymes. Suitable classes of enzymes include, but are not limited to, hydrolases such as proteases, carbohydrases, lipases; isomerases such as racemases, epimerases, tautomerases, or mutases; transferases, kinases, oxidoreductases, and phophatases. Suitable enzymes are listed in the Swiss-Prot enzyme database. Suitable protein backbones include, but are not limited to, all of those found in the protein data base compiled and serviced by the Research Collaboratory for Structural Bioinformatics (RCSB, formerly the Brookhaven National Lab).

Specifically, preferred scaffold proteins include, but are not limited to, those with known structures (including variants) including cytokines (IL-1ra (+receptor complex), IL-1 (receptor alone), IL-1a, IL-1b (including variants and or receptor complex), IL-2, IL-3, IL-4, IL-5, IL-6, IL-8, IL-10, IFN-β, INF-γ, IFN-α-2a; IFN-α-2B, TNF-α; CD40 ligand (chk), Human Obesity Protein Leptin, Granulocyte Colony-Stimulating Factor, Bone Morphogenetic Protein-7, Ciliary Neurotrophic Factor, Granulocyte-Macrophage Colony-Stimulating Factor, Monocyte Chemoattractant Protein 1, Macrophage Migration Inhibitory Factor, Human Glycosylation-Inhibiting Factor, Human Rantes, Human Macrophage Inflammatory Protein 1 Beta, human growth hormone, Leukemia Inhibitory Factor, Human Melanoma Growth Stimulatory Activity, neutrophil activating peptide-2, Cc-Chemokine Mcp-3, Platelet Factor M2, Neutrophil Activating Peptide 2, Eotaxin, Stromal Cell-Derived Factor-1, Insulin, Insulin-like Growth Factor I, Insulin-like Growth Factor II, Transforming Growth Factor B1, Transforming Growth Factor B2, Transforming Growth Factor B3, Transforming Growth Factor A, Vascular Endothelial growth factor (VEGF), acidic Fibroblast growth factor, basic Fibroblast growth factor, Endothelial growth factor, Nerve growth factor, Brain Derived Neurotrophic Factor, Ciliary Neurotrophic Factor, Platelet Derived Growth Factor, Human Hepatocyte Growth Factor, Glial Cell-Derived Neurotrophic Factor, (as well as the 55 cytokines in PDB 1/12/99)); Erythropoietin; other extracellular signalling moeities, including, but not limited to, hedgehog Sonic, hedgehog Desert, hedgehog Indian, hCG; coaguation factors including, but not limited to, TPA and Factor VIIa; transcription factors, including but not limited to, p53, p53 tetramerization domain, Zn fingers (of which more than 12 have structures), homeodomains (of which 8 have structures), leucine zippers (of which 4 have structures); antibodies,

including, but not limited to, cFv; viral proteins, including, but not limited to, hemagglutinin trimerization domain and hiv Gp41 ectodomain (fusion domain); intracellular signalling modules, including, but not limited to, SH2 domains (of which 8 structures are known), SH3 domains (of which 11 have structures), and Pleckstin Homology Domains; receptors, including, but not limited to, the extracellular Region Of Human Tissue Factor Cytokine-Binding Region Of Gp130, G-CSF receptor, erythropoietin receptor, Fibroblast Growth Factor receptor, TNF receptor, IL-1 receptor, IL-1 receptor/IL1ra complex, IL-4 receptor, INF-γ receptor alpha chain, MHC Class I, MHC Class II , T Cell Receptor, Insulin receptor, insulin receptor tyrosine kinase and human growth hormone receptor.

Once a scaffold protein is chosen, a primary library is generated using computational processing. Generally speaking, **in some embodiments,** the goal of the computational processing is to determine a set of optimized protein sequences. By "optimized protein sequence" herein is meant a sequence that best fits the mathematical equations of the computational process. As will be appreciated by those in the art, a global optimized sequence is the one sequence that best fits the equations (for example, when PDA is used, the global optimzed sequence is the sequence that best fits Equation 1, below); i.e. the sequence that has the lowest energy of any possible sequence. However, there are any number of sequences that are not the global minimum but that have low energies.

Thus, a "primary library" as used herein is a collection of optimized sequences, generally, but not always, in the form of a rank-ordered list. In theory, all possible sequences of a protein may be ranked; however, currently 10^{13} sequences is a practical limit. Thus, in general, some subset of all possible sequences is used as the primary library; generally, the top 10^3 to 10^{13} sequences are chosen as the primary library. The cutoff for inclusion in the rank ordered list of the primary library can be done in a variety of ways. For example, the cutoff may be just an arbitrary exclusion point: the top 10^5 sequences may comprise the primary library. Alternatively, all sequences scoring within a certain limit of the global optimum can be used; for example, all sequences with 10 kcal/mol of the global optimum could be used as the primary library. This method has the advantage of using a direct measure of fidelity to a three dimensional structure to determine inclusion. This approach can be used to insure that library mutations are not limited to positions that have the lowest energy gap between different mutations. Alternatively, the cutoff may be enforced when a predetermined number of mutations per position is reached. As a rank ordered sequence list is lengthened and the library is enlarged, more mutations per position are defined. Alternatively, the total number of sequences defined by the recombination of all mutations can be used as a cutoff criterion for the primary sequence library. Preferred values for the total number of sequences range from 100 to 10^{20}, particularly preferred values range from 1000 to 10^{13}, especially preferred values range from 1000 to 10^{7}. Alternatively, the first occurrence in the list of predefined undesirable residues can be used as a cutoff criterion. For example, the first hydrophilic residue occurring in a core position would limit the list. It should also be

noted that while these methods are described in conjunction with limiting the size of the primary library, these same techniques may be used to formulate the cutoff for inclusion in the secondary library as well.

Thus, the present invention provides methods to generate a primary library optionally comprising a rank ordered list of sequences, generally in terms of theoretical quantitative stability, as is more fully described below. Generating a primary library to optimize the stability of a conformation can be used to stabilize the active site transition state conformation of an enzyme, which will improve its activity. Similarly, stabilizing a ligand-receptor complex or enzyme-substrate complex will improve the binding affinity.

The primary libraries can be generated in a variety of ways. In essence, any methods that can result in either the relative ranking of the possible sequences of a protein based on measurable stability parameters, or a list of suitable sequences can be used. As will be appreciated by those in the art, any of the methods described herein or known in the art may be used alone, or in combination with other methods.

Generally, there are a variety of computational methods that can be used to generate a primary library. In a preferred embodiment, sequence based methods are used. Alternatively, structure based methods, such as PDA, described in detail below, are used.

In a preferred embodiment, the scaffold protein is an enzyme and highly accurate electrostatic models can be used for enzyme active site residue scoring to improve enzyme active site libraries (see Warshel, computer Modeling of Chemical Reactions in Enzymes and Solutions, Wiley & Sons, New York, (1991), hereby expressly incorporated by reference) These accurate models can assess the relative energies of sequences with high precision, but are computationally intensive.

Similarly, molecular dynamics calculations can be used to computationally screen sequences by individually calculating mutant sequence scores and compiling a rank ordered list.

In a preferred embodiment, residue pair potentials can be used to score sequences (Miyazawa et al., Macromolecules 18(3):534-552 (1985), expressly incorporated by reference) during computational screening.

In a preferred embodiment, sequence profile scores (Bowie et al., Science 253(5016):164-70 (1991), incorporated by reference) and/or potentials of mean force (Hendlich et al., J. Mol. Biol. 216(1):167-180 (1990), also incorporated by reference) can also be calculated to score sequences. These methods assess the match between a sequence and a 3D protein structure and hence can act to screen for fidelity to the protein structure. By using different scoring functions to rank sequences, different regions of sequence space can be sampled in the computational screen.

Furthermore, scoring functions can be used to screen for sequences that would create metal or co-factor binding sites in the protein (Hellinga, Fold Des. 3(1):R1-8 (1998), hereby expressly incorporated by reference). Similarly, scoring functions can be used to screen for sequences that would create disulfide bonds in the protein. These potentials attempt to specifically modify a protein structure to introduce a new structural motif.

In a preferred embodiment, sequence and/or structural alignment programs can be used to generate primary libraries. As is known in the art, there are a number of sequence-based alignment programs; including for example, Smith-Waterman searches, Needleman-Wunsch, Double Affine Smith-Waterman, frame search, Gribskov/GCG profile search, Gribskov/GCG profile scan, profile frame search, Bucher generalized profiles, Hidden Markov models, Hframe, Double Frame, Blast, Psi-Blast, Clustal, and GeneWise.

The source of the sequences can vary widely, and include taking sequences from one or more of the known databases, including, but not limited to, SCOP (Hubbard, et al., Nucleic Acids Res 27(1):254-256. (1999)); PFAM (Bateman, et al., Nucleic Acids Res 27(1):260-262. (1999)); VAST (Gibrat, et al., Curr Opin Struct Biol 6(3):377-385. (1996)); CATH (Orengo, et al., Structure 5(8):1093-1108. (1997)); PhD Predictor (http://www.embl-heidelberg.de/predictprotein /predictprotein.html); Prosite (Hofmann, et al., Nucleic Acids Res 27(1):215-219. (1999)); PIR (http://www.mips.biochem.mpg.de/proj/protseqdb/); GenBank (http://www.ncbi.nlm.nih.gov/); PDB (www.rcsb.org) and BIND (Bader, et al., Nucleic Acids Res 29(1):242-245. (2001)).

In addition, sequences from these databases can be subjected to continguous analysis or gene prediction; see Wheeler, et al., Nucleic Acids Res 28(1):10-14. (2000) and Burge and Karlin, J Mol Biol 268(1):78-94. (1997).

As is known in the art, there are a number of sequence alignment methodologies that can be used. For example, sequence homology based alignment methods can be used to create sequence

alignments of proteins related to the target structure (Altschul et al., J. Mol. Biol. 215(3):403 (1990), incorporated by reference). These sequence alignments are then examined to determine the observed sequence variations. These sequence variations are tabulated to define a primary library. In addition, as is further outlined below, these methods can also be used to generate secondary libraries.

Sequence based alignments can be used in a variety of ways. For example, a number of related proteins can be aligned, as is known in the art, and the "variable" and "conserved" residues defined; that is, the residues that vary or remain identical between the family members can be defined. These results can be used to generate a probability table, as outlined below. Similarly, these sequence variations can be tabulated and a secondary library defined from them as defined below. Alternatively, the allowed sequence variations can be used to define the amino acids considered at each position during the computational screening. Another variation is to bias the score for amino acids that occur in the sequence alignment, thereby increasing the likelihood that they are found during computational screening but still allowing consideration of other amino acids. This bias would result in a focused primary library but would not eliminate from consideration amino acids not found in the alignment. In addition, a number of other types of bias may be introduced. For example, diversity may be forced; that is, a "conserved" residue is chosen and altered to force diversity on the protein and thus sample a greater portion of the sequence space. Alternatively, the positions of high variability between family members (i.e. low conservation) can be randomized, either using all or a subset of amino acids. Similarly, outlier residues, either positional outliers or side chain outliers, may be eliminated.

Similarly, structural alignment of structurally related proteins can be done to generate sequence alignments. There are a wide variety of such structural alignment programs known. See for example VAST from the NCBI (http://www.ncbi.nlm.nih.gov:80/Structure/VAST/vast.shtml); SSAP (Orengo and Taylor, Methods Enzymol 266(617-635 (1996)) SARF2 (Alexandrov, Protein Eng 9(9):727-732. (1996)) CE (Shindyalov and Bourne, Protein Eng 11(9):739-747. (1998)); (Orengo et al., Structure 5(8):1093-108 (1997); Dali (Holm et al., Nucleic Acid Res. 26(1):316-9 (1998), all of which are incorporated by reference). These structurally-generated sequence alignments can then be examined to determine the observed sequence variations.

Primary libraries can be generated by predicting secondary structure from sequence, and then selecting sequences that are compatible with the predicted secondary structure. There are a number of secondary structure prediction methods, including, but not limited to, threading (Bryant and Altschul, Curr Opin Struct Biol 5(2):236-244. (1995)), Profile 3D (Bowie, et al., Methods Enzymol 266(598-616 (1996); MONSSTER (Skolnick, et al., J Mol Biol 265(2):217-241. (1997); Rosetta (Simons, et al., Proteins 37(S3):171-176 (1999); PSI-BLAST (Altschul and Koonin, Trends Biochem Sci 23(11):444-

447. (1998)); Impala (Schaffer, et al., Bioinformatics 15(12):1000-1011. (1999)); HMMER (McClure, et al., Proc Int Conf Intell Syst Mol Biol 4(155-164 (1996)); Clustal W (http://www.ebi.ac.uk/clustalw/); BLAST (Altschul, et al., J Mol Biol 215(3):403-410. (1990)), helix-coil transition theory (Munoz and Serrano, Biopolymers 41:495, 1997), neural networks, local structure alignment and others (e.g., see in Selbig et al., Bioinformatics 15:1039, 1999).

Similarly, as outlined above, other computational methods are known, including, but not limited to, sequence profiling (Bowie and Eisenberg, Science 253(5016): 164-70, (1991)), rotamer library selections (Dahiyat and Mayo, Protein Sci 5(5): 895-903 (1996); Dahiyat and Mayo, Science 278(5335): 82-7 (1997); Desjarlais and Handel, Protein Science 4: 2006-2018 (1995); Harbury et al, PNAS USA 92(18): 8408-8412 (1995); Kono et al., Proteins: Structure, Function and Genetics 19: 244-255 (1994); Hellinga and Richards, PNAS USA 91: 5803-5807 (1994)); and residue pair potentials (Jones, Protein Science 3: 567-574, (1994); PROSA (Heindlich et al., J. Mol. Biol. 216:167-180 (1990); THREADER (Jones et al., Nature 358:86-89 (1992), and other inverse folding methods such as those described by Simons et al. (Proteins, 34:535-543, 1999), Levitt and Gerstein (PNAS USA, 95:5913-5920, 1998), Godzik et al., PNAS, V89, PP 12098-102; Godzik and Skolnick (PNAS USA, 89:12098-102, 1992), Godzik et al. (J. Mol. Biol. 227:227-38, 1992) and two profile methods (Gribskov et al. PNAS 84:4355-4358 (1987) and Fischer and Eisenberg, Protein Sci. 5:947-955 (1996), Rice and Eisenberg J. Mol. Biol. 267:1026-1038(1997)), all of which are expressly incorporated by reference. In addition, other computational methods such as those described by Koehl and Levitt (J. Mol. Biol. 293:1161-1181 (1999); J. Mol. Biol. 293:1183-1193 (1999); expressly incorporated by reference) can be used to create a protein sequence library which can optionally then be used to generate a smaller secondary library for use in experimental screening for improved properties and function.

In addition, there are computational methods based on forcefield calculations such as SCMF that can be used as well for SCMF, see Delarue et la. Pac. Symp. Biocomput. 109-21 (1997), Koehl et al., J. Mol. Biol. 239:249 (1994); Koehl et al., Nat. Struc. Biol. 2:163 (1995); Koehl et al., Curr. Opin. Struct. Biol. 6:222 (1996); Koehl et al., J. Mol. Bio. 293:1183 (1999); Koehl et al., J. Mol. Biol. 293:1161 (1999); Lee J. Mol. Biol. 236:918 (1994); and Vasquez Biopolymers 36:53-70 (1995); all of which are expressly incorporated by reference. Other forcefield calculations that can be used to optimize the conformation of a sequence within a computational method, or to generate de novo optimized sequences as outlined herein include, but are not limited to, OPLS-AA (Jorgensen, et al., J. Am. Chem. Soc. (1996), v 118, pp 11225-11236; Jorgensen, W.L.; BOSS, Version 4.1; Yale University: New Haven, CT (1999)); OPLS (Jorgensen, et al., J. Am. Chem. Soc. (1988), v 110, pp 1657ff; Jorgensen, et al., J Am. Chem. Soc. (1990), v 112, pp 4768ff); UNRES (United Residue Forcefield; Liwo, et al., Protein Science (1993), v 2, pp1697-1714; Liwo, et al., Protein Science (1993), v 2, pp1715-1731; Liwo, et al., J. Comp. Chem. (1997), v 18, pp849-873; Liwo, et al., J. Comp. Chem.

(1997), v 18, pp874-884; Liwo, et al., J. Comp. Chem. (1998), v 19, pp259-276; Forcefield for Protein Structure Prediction (Liwo, et al., Proc. Natl. Acad. Sci. USA (1999), v 96, pp5482-5485); ECEPP/3 (Liwo et al., J Protein Chem 1994 May;13(4):375-80); AMBER 1.1 force field (Weiner, et al., J. Am. Chem. Soc. v106, pp765-784); AMBER 3.0 force field (U.C. Singh et al., Proc. Natl. Acad. Sci. USA. 82:755-759); CHARMM and CHARMM22 (Brooks, et al., J. Comp. Chem. v4, pp 187-217); cvff3.0 (Dauber-Osguthorpe, et al.,(1988) Proteins: Structure, Function and Genetics, v4,pp31-47); cff91 (Maple, et al., J. Comp. Chem. v15, 162-182); also, the DISCOVER (cvff and cff91) and AMBER forcefields are used in the INSIGHT molecular modeling package (Biosym/MSI, San Diego California) and HARMM is used in the QUANTA molecular modeling package (Biosym/MSI, San Diego California), all of which are expressly incorporated by reference. In fact, as is outlined below, these forcefield methods may be used to generate the secondary library directly; that is, no primary library is generated; rather, these methods can be used to generate a probability table from which the secondary library is directly generated, for example by using these forcefields during an SCMF calculation.

In a preferred embodiment, the computational method used to generate the primary library is Protein Design Automation (PDA), as is described in U.S.S.N.s 60/061,097, 60/043,464, 60/054,678, 09/127,926 and PCT US98/07254, all of which are expressly incorporated herein by reference. Briefly, PDA can be described as follows. A known protein structure is used as the starting point. The residues to be optimized are then identified, which may be the entire sequence or subset(s) thereof. The side chains of any positions to be varied are then removed. The resulting structure consisting of the protein backbone and the remaining sidechains is called the template. Each variable residue position is then preferably classified as a core residue, a surface residue, or a boundary residue; each classification defines a subset of possible amino acid residues for the position (for example, core residues generally will be selected from the set of hydrophobic residues, surface residues generally will be selected from the hydrophilic residues, and boundary residues may be either). Each amino acid can be represented by a discrete set of all allowed conformers of each side chain, called rotamers. Thus, to arrive at an optimal sequence for a backbone, all possible sequences of rotamers must be screened, where each backbone position can be occupied either by each amino acid in all its possible rotameric states, or a subset of amino acids, and thus a subset of rotamers.

Two sets of interactions are then calculated for each rotamer at every position: the interaction of the rotamer side chain with all or part of the backbone (the “singles” energy, also called the rotamer/template or rotamer/backbone energy), and the interaction of the rotamer side chain with all other possible rotamers at every other position or a subset of the other positions (the “doubles” energy, also called the rotamer/rotamer energy). The energy of each of these interactions is calculated through the use of a variety of scoring functions, which include the energy of van der

Waal's forces, the energy of hydrogen bonding, the energy of secondary structure propensity, the energy of surface area solvation and the electrostatics. Thus, the total energy of each rotamer interaction, both with the backbone and other rotamers, is calculated, and stored in a matrix form.

The discrete nature of rotamer sets allows a simple calculation of the number of rotamer sequences to be tested. A backbone of length n with m possible rotamers per position will have m^n possible rotamer sequences, a number which grows exponentially with sequence length and renders the calculations either unwieldy or impossible in real time. Accordingly, to solve this combinatorial search problem, a "Dead End Elimination" (DEE) calculation is performed. The DEE calculation is based on the fact that if the worst total interaction of a first rotamer is still better than the best total interaction of a second rotamer, then the second rotamer cannot be part of the global optimum solution. Since the energies of all rotamers have already been calculated, the DEE approach only requires sums over the sequence length to test and eliminate rotamers, which speeds up the calculations considerably. DEE can be rerun comparing pairs of rotamers, or combinations of rotamers, which will eventually result in the determination of a single sequence which represents the global optimum energy.

Once the global solution has been found, a Monte Carlo search may be done to generate a rank-ordered list of sequences in the neighborhood of the DEE solution. Starting at the DEE solution, random positions are changed to other rotamers, and the new sequence energy is calculated. If the new sequence meets the criteria for acceptance, it is used as a starting point for another jump. After a predetermined number of jumps, a rank-ordered list of sequences is generated. Monte Carlo searching is a sampling technique to explore sequence space around the global minimum or to find new local minima distant in sequence space. As is more additionally outlined below, there are other sampling techniques that can be used, including Boltzman sampling, genetic algorithm techniques and simulated annealing. In addition, for all the sampling techniques, the kinds of jumps allowed can be altered (e.g. random jumps to random residues, biased jumps (to or away from wild-type, for example), jumps to biased residues (to or away from similar residues, for example), etc.). Similarly, for all the sampling techniques, the acceptance criteria of whether a sampling jump is accepted can be altered.

As outlined in U.S.S.N. 09/127,926, the protein backbone (comprising (for a naturally occuring protein) the nitrogen, the carbonyl carbon, the α-carbon, and the carbonyl oxygen, along with the direction of the vector from the α-carbon to the β-carbon) may be altered prior to the computational analysis, by varying a set of parameters called supersecondary structure parameters.

Once a protein structure backbone is generated (with alterations, as outlined above) and input into the computer, explicit hydrogens are added if not included within the structure (for example, if the structure was generated by X-ray crystallography, hydrogens must be added). After hydrogen addition, energy minimization of the structure is run, to relax the hydrogens as well as the other atoms, bond angles and bond lengths. In a preferred embodiment, this is done by doing a number of steps of conjugate gradient minimization (Mayo *et al.*, J. Phys. Chem. 94:8897 (1990)) of atomic coordinate positions to minimize the Dreiding force field with no electrostatics. Generally from about 10 to about 250 steps is preferred, with about 50 being most preferred.

The protein backbone structure contains at least one variable residue position. As is known in the art, the residues, or amino acids, of proteins are generally sequentially numbered starting with the N-terminus of the protein. Thus a protein having a methionine at it's N-terminus is said to have a methionine at residue or amino acid position 1, with the next residues as 2, 3, 4, etc. At each position, the wild type (i.e. naturally occuring) protein may have one of at least 20 amino acids, in any number of rotamers. By "variable residue position" herein is meant an amino acid position of the protein to be designed that is not fixed in the design method as a specific residue or rotamer, generally the wild-type residue or rotamer.

In a preferred embodiment, all of the residue positions of the protein are variable. That is, every amino acid side chain may be altered in the methods of the present invention. This is particularly desirable for smaller proteins, although the present methods allow the design of larger proteins as well. While there is no theoretical limit to the length of the protein which may be designed this way, there is a practical computational limit.

In an alternate preferred embodiment, only some of the residue positions of the protein are variable, and the remainder are "fixed", that is, they are identified in the three dimensional structure as being in a set conformation. In some embodiments, a fixed position is left in its original conformation (which may or may not correlate to a specific rotamer of the rotamer library being used). Alternatively, residues may be fixed as a non-wild type residue; for example, when known site-directed mutagenesis techniques have shown that a particular residue is desirable (for example, to eliminate a proteolytic site or alter the substrate specificity of an enzyme), the residue may be fixed as a particular amino acid. Alternatively, the methods of the present invention may be used to evaluate mutations de novo, as is discussed below. In an alternate preferred embodiment, a fixed position may be "floated"; the amino acid at that position is fixed, but different rotamers of that amino acid are tested. In this embodiment, the variable residues may be at least one, or anywhere from 0.1% to 99.9% of the total number of residues. Thus, for example, it may be possible to change only a few (or one) residues, or most of the residues, with all possibilities in between.

In a preferred embodiment, residues which can be fixed include, but are not limited to, structurally or biologically functional residues; alternatively, biologically functional residues may specifically not be fixed. For example, residues which are known to be important for biological activity, such as the residues which form the active site of an enzyme, the substrate binding site of an enzyme, the binding site for a binding partner (ligand/receptor, antigen/antibody, etc.), phosphorylation or glycosylation sites which are crucial to biological function, or structurally important residues, such as disulfide bridges, metal binding sites, critical hydrogen bonding residues, residues critical for backbone conformation such as proline or glycine, residues critical for packing interactions, etc. may all be fixed in a conformation or as a single rotamer, or "floated".

Similarly, residues which may be chosen as variable residues may be those that confer undesirable biological attributes, such as susceptibility to proteolytic degradation, dimerization or aggregation sites, glycosylation sites which may lead to immune responses, unwanted binding activity, unwanted allostery, undesirable enzyme activity but with a preservation of binding, etc.

In a preferred embodiment, each variable position is classified as either a core, surface or boundary residue position, although in some cases, as explained below, the variable position may be set to glycine to minimize backbone strain. In addition, as outlined herein, residues need not be classified, they can be chosen as variable and any set of amino acids may be used. Any combination of core, surface and boundary positions can be utilized: core, surface and boundary residues; core and surface residues; core and boundary residues, and surface and boundary residues, as well as core residues alone, surface residues alone, or boundary residues alone.

The classification of residue positions as core, surface or boundary may be done in several ways, as will be appreciated by those in the art. In a preferred embodiment, the classification is done via a visual scan of the original protein backbone structure, including the side chains, and assigning a classification based on a subjective evaluation of one skilled in the art of protein modelling. Alternatively, a preferred embodiment utilizes an assessment of the orientation of the Cα-Cβ vectors relative to a solvent accessible surface computed using only the template Cα atoms, as outlined in U.S.S.N.s 60/061,097, 60/043,464, 60/054,678, 09/127,926 and PCT US98/07254. Alternatively, a surface area calculation can be done.

Once each variable position is classified as either core, surface or boundary, a set of amino acid side chains, and thus a set of rotamers, is assigned to each position. That is, the set of possible amino acid side chains that the program will allow to be considered at any particular position is chosen. Subsequently, once the possible amino acid side chains are chosen, the set of rotamers that will be

evaluated at a particular position can be determined. Thus, a core residue will generally be selected from the group of hydrophobic residues consisting of alanine, valine, isoleucine, leucine, phenylalanine, tyrosine, tryptophan, and methionine (in some embodiments, when the α scaling factor of the van der Waals scoring function, described below, is low, methionine is removed from the set), and the rotamer set for each core position potentially includes rotamers for these eight amino acid side chains (all the rotamers if a backbone independent library is used, and subsets if a rotamer dependent backbone is used). Similarly, surface positions are generally selected from the group of hydrophilic residues consisting of alanine, serine, threonine, aspartic acid, asparagine, glutamine, glutamic acid, arginine, lysine and histidine. The rotamer set for each surface position thus includes rotamers for these ten residues. Finally, boundary positions are generally chosen from alanine, serine, threonine, aspartic acid, asparagine, glutamine, glutamic acid, arginine, lysine histidine, valine, isoleucine, leucine, phenylalanine, tyrosine, tryptophan, and methionine. The rotamer set for each boundary position thus potentially includes every rotamer for these seventeen residues (assuming cysteine, glycine and proline are not used, although they can be). Additionally, in some preferred embodiments, a set of 18 naturally occuring amino acids (all except cysteine and proline, which are known to be particularly disruptive) are used.

Thus, as will be appreciated by those in the art, there is a computational benefit to classifying the residue positions, as it decreases the number of calculations. It should also be noted that there may be situations where the sets of core, boundary and surface residues are altered from those described above; for example, under some circumstances, one or more amino acids is either added or subtracted from the set of allowed amino acids. For example, some proteins which dimerize or multimerize, or have ligand binding sites, may contain hydrophobic surface residues, etc. In addition, residues that do not allow helix "capping" or the favorable interaction with an α-helix dipole may be subtracted from a set of allowed residues. This modification of amino acid groups is done on a residue by residue basis.

In a preferred embodiment, proline, cysteine and glycine are not included in the list of possible amino acid side chains, and thus the rotamers for these side chains are not used. However, in a preferred embodiment, when the variable residue position has a ϕ angle (that is, the dihedral angle defined by 1) the carbonyl carbon of the preceding amino acid; 2) the nitrogen atom of the current residue; 3) the α-carbon of the current residue; and 4) the carbonyl carbon of the current residue) greater than 0°, the position is set to glycine to minimize backbone strain.

Once the group of potential rotamers is assigned for each variable residue position, processing proceeds as outlined in U.S.S.N. 09/127,926 and PCT US98/07254. This processing step entails analyzing interactions of the rotamers with each other and with the protein backbone to generate

optimized protein sequences. Simplistically, the processing initially comprises the use of a number of scoring functions to calculate energies of interactions of the rotamers, either to the backbone itself or other rotamers. Preferred PDA scoring functions include, but are not limited to, a Van der Waals potential scoring function, a hydrogen bond potential scoring function, an atomic solvation scoring function, a secondary structure propensity scoring function and an electrostatic scoring function. As is further described below, at least one scoring function is used to score each position, although the scoring functions may differ depending on the position classification or other considerations, like favorable interaction with an α-helix dipole. As outlined below, the total energy which is used in the calculations is the sum of the energy of each scoring function used at a particular position, as is generally shown in Equation 1:

Equation 1

$$E_{total} = nE_{vdw} + nE_{as} + nE_{h\text{-}bonding} + nE_{ss} + nE_{elec}$$

In Equation 1, the total energy is the sum of the energy of the van der Waals potential (E_{vdw}), the energy of atomic solvation (E_{as}), the energy of hydrogen bonding ($E_{h\text{-}bonding}$), the energy of secondary structure (E_{ss}) and the energy of electrostatic interaction (E_{elec}). The term n is either 0 or 1, depending on whether the term is to be considered for the particular residue position.

As outlined in U.S.S.N.s 60/061,097, 60/043,464, 60/054,678, 09/127,926 and PCT US98/07254, any combination of these scoring functions, either alone or in combination, may be used. Once the scoring functions to be used are identified for each variable position, the preferred first step in the computational analysis comprises the determination of the interaction of each possible rotamer with all or part of the remainder of the protein. That is, the energy of interaction, as measured by one or more of the scoring functions, of each possible rotamer at each variable residue position with either the backbone or other rotamers, is calculated. In a preferred embodiment, the interaction of each rotamer with the entire remainder of the protein, i.e. both the entire template and all other rotamers, is done. However, as outlined above, it is possible to only model a portion of a protein, for example a domain of a larger protein, and thus in some cases, not all of the protein need be considered. The term "portion", as used herein, with regard to a protein refers to a fragment of that protein. This fragment may range in size from 10 amino acid residues to the entire amino acid sequence minus one amino acid. Accordingly, the term "portion", as used herein, with regard to a nucleic refers to a fragment of that nucleic acid. This fragment may range in size from 10 nucleotides to the entire nucleic acid sequence minus one nucleotide.

In a preferred embodiment, the first step of the computational processing is done by calculating two sets of interactions for each rotamer at every position: the interaction of the rotamer side chain with the template or backbone (the "singles" energy), and the interaction of the rotamer side chain with all

other possible rotamers at every other position (the "doubles" energy), whether that position is varied or floated. It should be understood that the backbone in this case includes both the atoms of the protein structure backbone, as well as the atoms of any fixed residues, wherein the fixed residues are defined as a particular conformation of an amino acid.

Thus, "singles" (rotamer/template) energies are calculated for the interaction of every possible rotamer at every variable residue position with the backbone, using some or all of the scoring functions. Thus, for the hydrogen bonding scoring function, every hydrogen bonding atom of the rotamer and every hydrogen bonding atom of the backbone is evaluated, and the E_{HB} is calculated for each possible rotamer at every variable position. Similarly, for the van der Waals scoring function, every atom of the rotamer is compared to every atom of the template (generally excluding the backbone atoms of its own residue), and the E_{vdW} is calculated for each possible rotamer at every variable residue position. In addition, generally no van der Waals energy is calculated if the atoms are connected by three bonds or less. For the atomic solvation scoring function, the surface of the rotamer is measured against the surface of the template, and the E_{as} for each possible rotamer at every variable residue position is calculated. The secondary structure propensity scoring function is also considered as a singles energy, and thus the total singles energy may contain an E_{ss} term. As will be appreciated by those in the art, many of these energy terms will be close to zero, depending on the physical distance between the rotamer and the template position; that is, the farther apart the two moieties, the lower the energy.

For the calculation of "doubles" energy (rotamer/rotamer), the interaction energy of each possible rotamer is compared with every possible rotamer at all other variable residue positions. Thus, "doubles" energies are calculated for the interaction of every possible rotamer at every variable residue position with every possible rotamer at every other variable residue position, using some or all of the scoring functions. Thus, for the hydrogen bonding scoring function, every hydrogen bonding atom of the first rotamer and every hydrogen bonding atom of every possible second rotamer is evaluated, and the E_{HB} is calculated for each possible rotamer pair for any two variable positions. Similarly, for the van der Waals scoring function, every atom of the first rotamer is compared to every atom of every possible second rotamer, and the E_{vdW} is calculated for each possible rotamer pair at every two variable residue positions. For the atomic solvation scoring function, the surface of the first rotamer is measured against the surface of every possible second rotamer, and the E_{as} for each possible rotamer pair at every two variable residue positions is calculated. The secondary structure propensity scoring function need not be run as a "doubles" energy, as it is considered as a component of the "singles" energy. As will be appreciated by those in the art, many of these double energy terms will be close to zero, depending on the physical distance between the first rotamer and the second rotamer; that is, the farther apart the two moieties, the lower the energy.

In addition, as will be appreciated by those in the art, a variety of force fields that can be used in the PCA calculations can be used, including, but not limited to, Dreiding I and Dreiding II (Mayo et al, J. Phys. Chem. 948897 (1990)), AMBER (Weiner et al., J. Amer. Chem. Soc. 106:765 (1984) and Weiner et al., J. Comp. Chem. 106:230 (1986)), MM2 (Allinger J. Chem. Soc. 99:8127 (1977), Liljefors et al., J. Com. Chem. 8:1051 (1987)); MMP2 (Sprague et al., J. Comp. Chem. 8:581 (1987)); CHARMM (Brooks et al., J. Comp. Chem. 106:187 (1983)); GROMOS; and MM3 (Allinger et al., J. Amer. Chem. Soc. 111:8551 (1989)), OPLS-AA (Jorgensen, et al., J. Am. Chem. Soc. (1996), v 118, pp 11225-11236; Jorgensen, W.L.; BOSS, Version 4.1; Yale University: New Haven, CT (1999)); OPLS (Jorgensen, et al., J. Am. Chem. Soc. (1988), v 110, pp 1657ff; Jorgensen, et al., J Am. Chem. Soc. (1990), v 112, pp 4768ff); UNRES (United Residue Forcefield; Liwo, et al., Protein Science (1993), v 2, pp1697-1714; Liwo, et al., Protein Science (1993), v 2, pp1715-1731; Liwo, et al., J. Comp. Chem. (1997), v 18, pp849-873; Liwo, et al., J. Comp. Chem. (1997), v 18, pp874-884; Liwo, et al., J. Comp. Chem. (1998), v 19, pp259-276; Forcefield for Protein Structure Prediction (Liwo, et al., Proc. Natl. Acad. Sci. USA (1999), v 96, pp5482-5485); ECEPP/3 (Liwo et al., J Protein Chem 1994 May;13(4):375-80); AMBER 1.1 force field (Weiner, et al., J. Am. Chem. Soc. v106, pp765-784); AMBER 3.0 force field (U.C. Singh et al., Proc. Natl. Acad. Sci. USA. 82:755-759); CHARMM and CHARMM22 (Brooks, et al., J. Comp. Chem. v4, pp 187-217); cvff3.0 (Dauber-Osguthorpe, et al.,(1988) Proteins: Structure, Function and Genetics, v4,pp31-47); cff91 (Maple, et al., J. Comp. Chem. v15, 162-182); also, the DISCOVER (cvff and cff91) and AMBER forcefields are used in the INSIGHT molecular modeling package (Biosym/MSI, San Diego California) and HARMM is used in the QUANTA molecular modeling package (Biosym/MSI, San Diego California), all of which are expressly incorporated by reference.

Once the singles and doubles energies are calculated and stored, the next step of the computational processing may occur. As outlined in U.S.S.N. 09/127,926 and PCT US98/07254, preferred embodiments utilize a Dead End Elimination (DEE) step, and preferably a Monte Carlo step.

PDA, viewed broadly, has three components that may be varied to alter the output (e.g. the primary library): the scoring functions used in the process; the filtering technique, and the sampling technique.

In a preferred embodiment, the scoring functions may be altered. In a preferred embodiment, the scoring functions outlined above may be biased or weighted in a variety of ways. For example, a bias towards or away from a reference sequence or family of sequences can be done; for example, a bias towards wild-type or homolog residues may be used. Similarly, the entire protein or a fragment of it may be biased; for example, the active site may be biased towards wild-type residues, or domain residues towards a particular desired physical property can be done. Furthermore, a bias towards or against increased energy can be generated. Additional scoring function biases include, but are not

limited to applying electrostatic potential gradients or hydrophobicity gradients, adding a substrate or binding partner to the calculation, or biasing towards a desired charge or hydrophobicity.

In addition, in an alternative embodiment, there are a variety of additional scoring functions that may be used. Additional scoring functions include, but are not limited to torsional potentials, or residue pair potentials, or residue entropy potentials. Such additional scoring functions can be used alone, or as functions for processing the library after it is scored initially. For example, a variety of functions derived from data on binding of peptides to MHC (Major Histocompatibility Complex) can be used to rescore a library in order to eliminate proteins containing sequences which can potentially bind to MHC, i.e. potentially immunogenic sequences.

In a preferred embodiment, a variety of filtering techniques can be done, including, but not limited to, DEE and its related counterparts. Additional filtering techniques include, but are not limited to branch-and-bound techniques for finding optimal sequences (Gordon and Majo, Structure Fold. Des. 7:1089-98, 1999), and exhaustive enumeration of sequences. It should be noted however, that some techniques may also be done without any filtering techniques; for example, sampling techniques can be used to find good sequences, in the absence of filtering.

As will be appreciated by those in the art, once an optimized sequence or set of sequences is generated, **(or again, these need not be optimized or ordered)** a variety of sequence space sampling methods can be done, either in addition to the preferred Monte Carlo methods, or instead of a Monte Carlo search. That is, once a sequence or set of sequences is generated, preferred methods utilize sampling techniques to allow the generation of additional, related sequences for testing.

These sampling methods can include the use of amino acid substitutions, insertions or deletions, or recombinations of one or more sequences. As outlined herein, a preferred embodiment utilizes a Monte Carlo search, which is a series of biased, systematic, or random jumps. However, there are other sampling techniques that can be used, including Boltzman sampling, genetic algorithm techniques and simulated annealing. In addition, for all the sampling techniques, the kinds of jumps allowed can be altered (e.g. random jumps to random residues, biased jumps (to or away from wild-type, for example), jumps to biased residues (to or away from similar residues, for example), etc.). Jumps where multiple residue positions are coupled (two residues always change together, or never change together), jumps where whole sets of residues change to other sequences (e.g., recombination). Similarly, for all the sampling techniques, the acceptance criteria of whether a sampling jump is accepted can be altered, to allow broad searches at high temperature and narrow

searches close to local optima at low temperatures. See Metropolis et al., J. Chem Phys v21, pp 1087, 1953, hereby expressly incorporated by reference.

In addition, it should be noted that the preferred methods of the invention result in a rank ordered list of sequences; that is, the sequences are ranked on the basis of some objective criteria. However, as outlined herein, it is possible to create a set of non-ordered sequences, for example by generating a probability table directly (for example using SCMF analysis or sequence alignment techniques) that lists sequences without ranking them. The sampling techniques outlined herein can be used in either situation.

In a preferred embodiment, Boltzman sampling is done. As will be appreciated by those in the art, the temperature criteria for Boltzman sampling can be altered to allow broad searches at high temperature and narrow searches close to local optima at low temperatures (see e.g., Metropolis et al., J. Chem. Phys. 21:1087, 1953).

In a preferred embodiment, the sampling technique utilizes genetic algorithms, e.g., such as those described by Holland (Adaptation in Natural and Artifical Systems, 1975, Ann Arbor, U. Michigan Press). Genetic algorithm analysis generally takes generated sequences and recombines them computationally, similar to a nucleic acid recombination event, in a manner similar to "gene shuffling". Thus the "jumps" of genetic algorithm analysis generally are multiple position jumps. In addition, as outlined below, correlated multiple jumps may also be done. Such jumps can occur withdifferent crossover positions and more than one recombination at a time, and can involve recombination of two or more sequences. Furthermore, deletions or insertions (random or biased) can be done. In addition, as outlined below, genetic algorithm analysis may also be used after the secondary library has been generated.

In a preferred embodiment, the sampling technique utilizes simulated annealing, e.g., such as described by Kirkpatrick et al. (Science, 220:671-680, 1983). Simulated annealing alters the cutoff for accepting good or bad jumps by altering the temperature. That is, the stringency of the cutoff is altered by altering the temperature. This allows broad searches at high temperature to new areas of sequence space, altering with narrow searches at low temperature to explore regions in detail.

In addition, as outlined below, these sampling methods can be used to further process a secondary library to generate additional secondary libraries (sometimes referred to herein as tertiary libraries).

Thus, the primary library can be generated in a variety of computational ways, including structure based methods such as PDA, or sequence based methods, or combinations as outlined herein.

Accordingly, the computational processing **results in a set of sequences, that may be optimized protein sequences if some sort of ranking or scoring functions are used.** These optimized protein sequences are generally, but not always, significantly different from the wild-type sequence from which the backbone was taken. That is, each optimized protein sequence preferably comprises at least about 5-10% variant amino acids from the starting or wild-type sequence, with at least about 15-20% changes being preferred and at least about 30% changes being particularly preferred.

The cutoff for the primary library is then enforced, resulting in a set of primary sequences forming the primary library. As outlined above, this may be done in a variety of ways, including an arbitrary cutoff, an energy limitation, or when a certain number of residue positions have been varied. In general, the size of the primary library will vary with the size of the protein, the number of residues that are changing, the computational methods used, the cutoff applied and the discretion of the user. In general, it is preferable to have the primary library be large enough to randomly sample a reasonable sequence space to allow for robust secondary libraries. Thus, primary libraries that range from about 50 to about 10^{13} are preferred, with from about 1000 to about 10^{7} being particularly preferred, and from about 1000 to about 100,000 being especially preferred.

In a preferred embodiment **when scoring is used**, although this is not required, the primary library comprises the globally optimal sequence in its optimal conformation, i.e. the optimum rotamer at each variable position. That is, computational processing is run until the simulation program converges on a single sequence which is the global optimum. In a preferred embodiment, the primary library comprises at least two optimized protein sequences. Thus for example, the computational processing step may eliminate a number of disfavored combinations but be stopped prior to convergence, providing a library of sequences of which the global optimum is one. In addition, further computational analysis, for example using a different method, may be run on the library, to further eliminate sequences or rank them differently. Alternatively, as is more fully described in U.S.S.N.s 60/061,097, 60/043,464, 60/054,678, 09/127,926 and PCT US98/07254, the global optimum may be reached, and then further computational processing may occur, which generates additional optimized sequences in the neighborhood of the global optimum.

In addition, in some embodiments, primary library sequences that did not make the cutoff are included in the primary library. This may be desirable in some situations to evaluate the primary library

generation method, to serve as controls or comparisons, or to sample additional sequence space. For example, in a preferred embodiment, the wild-type sequence is included.

It should also be noted that different ranking systems can be used. For example, a list of naturally occurring sequences can be used to calculate all possible recombinations of these sequences, with an optional rank ordering step. Alternatively, once a primary library is generated, one could rank order only those recombinations that occur at cross-over points with at least a threshold of identity over a given window. For example, 100% identity over a window of 6 amino acids, or 80% identity over a window of 10 amino acids. Alternatively, as for all the systems outlined herein, the homology could be considered at the DNA level, by computationally considering the translation fo the amino acids to their respective DNA codons. Different codon usages could be considered. A preferred embodiment considers only recombinations with corssover points that have DNA sequence identity sufficient for DNA hybridization of the differing sequences.

As is further outlined below, It should also be noted that combining different primary libraries may be done. For example, positions in a protein that show a great deal of mutational diversity in computational screening can be fixed as outlined below and a different primary library regenerated. A rank ordered list of the same length as the first would now show diversity in previously rarely changing positions. The variants from the first primary library can be combined with the variants from the second primary library to provide a combined library at lower computational cost than creating a very long rank ordered list. This approach can be particularly useful to sample sequence diversity in both low energy gap, readily changing surface positions and high energy gap, rarely changing core positions. **In addition, primary libraries can be generated by combining one or more of the different calculations to form one big primary library.**

Thus, the present invention provides primary libraries comprising a list of computationally derived sequences. In a preferred embodiment, these sequences are in the form of a rank ordered list. From this primary library, a secondary library is generated. As outlined herein, there are a number of different ways to generate a secondary library.

In a preferred embodiment, the primary library of the scaffold protein is used to generate a secondary library. As will be appreciated by those in the art, the secondary library can be either a subset of the primary library, or contain new library members, i.e. sequences that are not found in the primary library. That is, in general, the variant positions and/or amino acid residues in the variant positions can be recombined in any number of ways to form a new library that exploits the sequence variations found in the primary library. That is, having identified "hot spots" or important variant positions and/or

residues, these positions can be recombined in novel ways to generate novel sequences to form a secondary library. Thus, in a preferred embodiment, the secondary library comprises at least one member sequence that is not found in the primary library, and preferably a plurality of such sequences.

In one embodiment, all or a portion of the primary library serves as the secondary library. That is, a cutoff is applied to the primary sequences and these sequences serve as the secondary library, without further manipulation or recombination. The library members can be made as outlined below, e.g. by direct synthesis or by constructing the nucleic acids encoding the library members, expressing them in a suitable host, optionally followed by screening.

In a preferred embodiment, the secondary library is generated by tabulating the amino acid positions that vary from a reference sequence. The reference sequence can be arbitrarily selected, or preferably is chosen either as the wild-type sequence or the global optimum sequence, with the latter being preferred. That is, each amino acid position that varies in the primary library is tabulated. Of course, if the original computational analysis fixed some positions, the variable positions of the secondary library will comprise either just these original variable positions or some subset of these original variable positions. That is, assuming a protein of 100 amino acids, the original computational screen can allow all 100 positions to be varied. However, due to the cutoff in the primary library, only 25 positions may vary. Alternatively, assuming the same 100 amino acid protein, the original computational screen could have varied only 25 positions, keeping the other 75 fixed; this could result in only 12 of the 25 being varied in the cutoff primary library. These primary library positions can then be recombined to form a secondary library, wherein all possible combinations of these variable positions form the secondary library. It should be noted that the non-variable positions are set to the reference sequence positions.

The formation of the secondary library using this method may be done in two general ways; either all variable positions are allowed to be any amino acid, or subsets of amino acids are allowed for each position.

In a preferred embodiment, all amino acid residues are allowed at each variable position identified in the primary library. That is, once the variable positions are identified, a secondary library comprising every combination of every amino acid at each variable position is made.

In a preferred embodiment, subsets of amino acids are chosen. The subset at any position may be either chosen by the user, or may be a collection of the amino acid residues generated in the primary screen. That is, assuming core residue 25 is variable and the primary screen gives 5 different possible amino acids for this position, the user may chose the set of good core residues outlined above (e.g. hydrophobic residues), or the user may build the set by chosing the 5 different amino acids generated in the primary screen. Alternatively, combinations of these techniques may be used, wherein the set of identified residues is manually expanded. For example, in some embodiments, fewer than the number of amino acid residues is chosen; for example, only three of the five may be chosen. Alternatively, the set is manually expanded; for example, if the computation picks two different hydrophobic residues, additional choices may be added. Similarly, the set may be biased, for example either towards or away from the wild-type sequence, or towards or away from known domains, etc.

In addition, this may be done by analyzing the primary library to determine which amino acid positions in the scaffold protein have a high mutational frquency, and which positions have a low mutation frequency. The secondary library can be generated by randomizing the amino acids at the positions that have high numbers of mutations, while keeping constant the positions that do not have mutations above a certain frequency. For example, if the position has less than 20% and more preferably 10% mutations, it may be kept constant as the reference sequence position.

In a preferred embodiment, the secondary library is generated from a probability distribution table. As outlined herein, there are a variety of methods of generating a probability distribution table, including using PDA, sequence alignments, forcefield calculations such as SCMF calculations, etc. In addition, the probability distribution can be used to generate information entropy scores for each position, as a measure of the mutational frequency observed in the library.

In this embodiment, the frequency of each amino acid residue at each variable position in the list is identified. Frequencies can be thresholded, wherein any variant frequency lower than a cutoff is set to zero. This cutoff is preferably 1%, 2%, 5%, 10% or 20%, with 10% being particularly preferred. These frequencies are then built into the secondary library. That is, as above, these variable positions are collected and all possible combinations are generated, but the amino acid residues that "fill" the secondary library are utilized on a frequency basis. Thus, in a non-frequency based secondary library, a variable position that has 5 possible residues will have 20% of the proteins comprising that variable position with the first possible residue, 20% with the second, etc. However, in a frequency based secondary library, a variable position that has 5 possible residues with frequencies of 10%, 15%, 25%, 30% and 20%, respectively, will have 10% of the proteins comprising that variable position with the first possible residue, 15% of the proteins with the second residue, 25% with the third, etc. As will be

appreciated by those in the art, the actual frequency may depend on the method used to actually generate the proteins; for example, exact frequencies may be possible when the proteins are synthesized. However, when the frequency-based primer system outlined below is used, the actual frequencies at each position will vary, as outlined below.

As will be appreciated by those in the art and outlined herein, probability distribution tables can be generated in a variety of ways. In addition to the methods outlined herein, self-consistent mean field (SCMF) methods can be used in the direct generation of probability tables. SCMF is a deterministic computational method that uses a mean field description of rotamer interactions to calculate energies. A probability table generated in this way can be used to create secondary libraries as described herein. SCMF can be used in three ways: the frequencies of amino acids and rotamers for each amino acid are listed at each position; the probabilities are determined directly from SCMF (see Delarue et la. Pac. Symp. Biocomput. 109-21 (1997), expressly incorporated by reference). In addition, highly variable positions and non-variable positions can be identified. Alternatively, another method is used to determine what sequence is jumped to during a search of sequence space; SCMF is used to obtain an accurate energy for that sequence; this energy is then used to rank it and create a rank-ordered list of sequences (similar to a Monte Carlo sequence list). A probability table showing the frequencies of amino acids at each position can then be calculated from this list (Koehl et al., J. Mol. Biol. 239:249 (1994); Koehl et al., Nat. Struc. Biol. 2:163 (1995); Koehl et al., Curr. Opin. Struct. Biol. 6:222 (1996); Koehl et al., J. Mol. Bio. 293:1183 (1999); Koehl et al., J. Mol. Biol. 293:1161 (1999); Lee J. Mol. Biol. 236:918 (1994); and Vasquez Biopolymers 36:53-70 (1995); all of which are expressly incorporated by reference. Similar methods include, but are not limited to, OPLS-AA (Jorgensen, et al., J. Am. Chem. Soc. (1996), v 118, pp 11225-11236; Jorgensen, W.L.; BOSS, Version 4.1; Yale University: New Haven, CT (1999)); OPLS (Jorgensen, et al., J. Am. Chem. Soc. (1988), v 110, pp 1657ff; Jorgensen, et al., J Am. Chem. Soc. (1990), v 112, pp 4768ff); UNRES (United Residue Forcefield; Liwo, et al., Protein Science (1993), v 2, pp1697-1714; Liwo, et al., Protein Science (1993), v 2, pp1715-1731; Liwo, et al., J. Comp. Chem. (1997), v 18, pp849-873; Liwo, et al., J. Comp. Chem. (1997), v 18, pp874-884; Liwo, et al., J. Comp. Chem. (1998), v 19, pp259-276; Forcefield for Protein Structure Prediction (Liwo, et al., Proc. Natl. Acad. Sci. USA (1999), v 96, pp5482-5485); ECEPP/3 (Liwo et al., J Protein Chem 1994 May;13(4):375-80); AMBER 1.1 force field (Weiner, et al., J. Am. Chem. Soc. v106, pp765-784); AMBER 3.0 force field (U.C. Singh et al., Proc. Natl. Acad. Sci. USA. 82:755-759); CHARMM and CHARMM22 (Brooks, et al., J. Comp. Chem. v4, pp 187-217); cvff3.0 (Dauber-Osguthorpe, et al.,(1988) Proteins: Structure, Function and Genetics, v4,pp31-47); cff91 (Maple, et al., J. Comp. Chem. v15, 162-182); also, the DISCOVER (cvff and cff91) and AMBER forcefields are used in the INSIGHT molecular modeling package (Biosym/MSI, San Diego California) and HARMM is used in the QUANTA molecular modeling package (Biosym/MSI, San Diego California).

In addition, as outlined herein, a preferred method of generating a probability distribution table is through the use of sequence alignment programs. In addition, the probability table can be obtained by a combination of sequence alignments and computational approaches. For example, one can add amino acids found in the alignment of homologous sequences to the result of the computation. Preferable one can add the wild type amino acid identity to the probability table if it is not found in the computation.

As will be appreciated, a secondary library created by recombining variable positions and/or residues at the variable position may not be in a rank-ordered list. In some embodiments, the entire list may just be made and tested. Alternatively, in a preferred embodiment, the secondary library is also in the form of a rank ordered list. This may be done for several reasons, including the size of the secondary library is still too big to generate experimentally, or for predictive purposes. This may be done in several ways. In one embodiment, the secondary library is ranked using the scoring functions of PDA to rank the library members. Alternatively, statistical methods could be used. For example, the secondary library may be ranked by frequency score; that is, proteins containing the most of high frequency residues could be ranked higher, etc. This may be done by adding or multiplying the frequency at each variable position to generate a numerical score. Similarly, the secondary library different positions could be weighted and then the proteins scored; for example, those containing certain residues could be arbitrarily ranked.

As outlined herein, secondary libraries can be generated in two general ways. The first is computationally, as above, wherein the primary library is further computationally manipulated, for example by recombining the possible variant positions and/or amino acid residues at each variant position or by recombining portions of the sequences containing one or more variant position. It may be ranked, as outlined above. This computationally-derived secondary library can then be experimentally generated by synthesizing the library members or nucleic acids encoding them, as is more fully outlined below. Alternatively, the secondary library is made experimentally; that is, nucleic acid recombination techniques are used to experimentally generate the combinations. This can be done in a variety of ways, as outlined below.

In a preferred embodiment, the different protein members of the secondary library may be chemically synthesized. This is particularly useful when the designed proteins are short, preferably less than 150 amino acids in length, with less than 100 amino acids being preferred, and less than 50 amino acids being particularly preferred, although as is known in the art, longer proteins can be made chemically or enzymatically. See for example Wilken et al, Curr. Opin. Biotechnol. 9:412-26 (1998), hereby expressly incorporated by reference.

In a preferred embodiment, particularly for longer proteins or proteins for which large samples are desired, the secondary library sequences are used to create nucleic acids such as DNA which encode the member sequences and which can then be cloned into host cells, expressed and assayed, if desired. Thus, nucleic acids, and particularly DNA, can be made which encodes each member protein sequence. This is done using well known procedures. The choice of codons, suitable expression vectors and suitable host cells will vary depending on a number of factors, and can be easily optimized as needed.

In a preferred embodiment, multiple PCR reactions with pooled oligonucleotides is done, as is generally depicted in Figure 1. In this embodiment, overlapping oligonucleotides are synthesized which correspond to the full length gene. Again, these oligonucleotides may represent all of the different amino acids at each variant position or subsets.

In a preferred embodiment, these oligonucleotides are pooled in equal proportions and multiple PCR reactions are performed to create full length sequences containing the combinations of mutations defined by the secondary library. In addition, this may be done using error-prone PCR methods.

In a preferred embodiment, the different oligonucleotides are added in relative amounts corresponding to the probability distribution table. The multiple PCR reactions thus result in full length sequences with the desired combinations of mutaions in the desired proportions.

The total number of oligonucleotides needed is a function of the number of positions being mutated and the number of mutations being considered at these positions:

(number of oligos for constant positions) + M1 + M2 + M3 +... Mn = (total number of oligos required), where Mn is the number of mutations considered at position n in the sequence.

In a preferred embodiment, each overlapping oligonucleotide comprises only one position to be varied; in alternate embodiments, the variant positions are too close together to allow this and multiple variants per oligonucleotide are used to allow complete recombination of all the possibilities. That is, each oligo can contain the codon for a single position being mutated, or for more than one position being mutated. The multiple positions being mutated must be close in sequence to prevent the oligo length from being impractical. For multiple mutating positions on an oligonucleotide, particular combinations of mutations can be included or excluded in the library by including or excluding the oligonucleotide encoding that combination. For example, as discussed herein, there may be

correlations between variable regions; that is, when position X is a certain residue, position Y must (or must not) be a particular residue. These sets of variable positions are sometimes referred to herein as a "cluster". When the clusters are comprised of residues close together, and thus can reside on one oligonuclotide primer, the clusters can be set to the "good" correlations, and eliminate the bad combinations that may decrease the effectiveness of the library. However, if the residues of the cluster are far apart in sequence, and thus will reside on different oligonuclotides for synthesis, it may be desirable to either set the residues to the "god" correlation, or eliminate them as variable residues entirely. In an alternative embodiment,the library may be generated in several steps, so that the cluster mutations only appear together. This procedure, i.e., the procedure of identifying mutation clusters and either placing them on the same oligonucleotides oreliminating them from the library or library generation in several steps preserving clusters, can considerably enrich the experimental library with properly folded protein. Identification of clusters can be carried out by a number of wasy, e.g. by using known pattern recognition methods, comparisons of frequencies of occurrence of mutations or by using energy analysis of the sequences to be experimentally generated (for example, if the energy of interaction is high, the positions are correlated). these correlations may be positional correlations (e.g. variable positions 1 and 2 always change together or never change together) or sequence correlations (e.g. if there is a residue A at position 1, there is always residue B at position 2). See: Pattern discovery in Biomolecular Data: Tools, Techniques, and Applications; edited by Jason T.L. Wang, Bruce A. Shapiro, Dennis Shasha. New York: Oxford Unviersity, 1999; Andrews, Harry C. Introduction to mathematical techniques in patter recognition; New York, Wiley-Interscience [1972]; Applications of Pattern Recognition; Editor, K.S. Fu. Boca Raton, Fla. CRC Press, 1982; Genetic Algorithms for Pattern Recognition; edited by Sankar K. Pal, Paul P. Wang. Boca Raton : CRC Press, c1996; Pandya, Abhijit S., Pattern recognition with Neural networks in C++/Abhijit S. Pandya, Robert B. Macy. Boca Raton, Fla.: CRC Press, 1996; Handbook of pattern recognition and computer vision / edited by C.H. Chen, L.F. Pau, P.S.P. Wang. 2nd ed. Signapore ; River Edge, N.J. : World Scientific, c1999; Friedman, Introduction to Pattern Recognition : Statistical, Structural, Neural, and Fuzzy Logic Approaches ; River Edge, N.J. : World Scientific, c1999, Series title: Serien a machine perception and artificial intelligence; vol. 32; all of wh ich are expressly incorporated by reference. In addition programs used to search for consensus motifs can be used as well.

In addition, correlations and shuffling can be fixed or optimized by altering the design of the oligonucleotides; that is, by deciding where the oligonucleotides (primers) start and stop (e.g. where the sequences are "cut"). The start and stop sites of oligos can be set to maximize the number of clusters that appear in single oligonucleotides, thereby enriching the library with higher scoring sequences. Different oligonucleotides start and stop site options can be computationally modeled and ranked according to number of clusters that are represented on single oligos, or the percentage of the resulting sequences consistent with the predicted libarary of sequences.

The total number of oligonucleotides required increases when multiple mutable positions are encoded by a single oligonucleotide. The annealed regions are the ones that remain constant, i.e. have the sequence of the reference sequence.

Oligonucleotides with insertions or deletions of codons can be used to create a library expressing different length proteins. In particular computational sequence screening for insertions or deletions can result in secondary libraries defining different length proteins, which can be expressed by a library of pooled oligonucleotide of different lengths.

In a preferred embodiment, the secondary library is done by shuffling the famil7 (e.g. a set of variants); that is, some set of the top sequences (if a rank-ordered list is used) can be shuffled, either with or without error-prone PCR. "Shuffling" in this context means a recombination of related sequences, generally in a random way. It can include "shuffling" as defined and exemplified in U.S. Patent Nso. 5,830,721; 5,811,238; 5,605,793; 5,837,458 and PCT US/19256, all of which are expressly incorporated by reference in their entirety. This set of sequences can also be an artificial set; for example, from a probability table (for example generated using SCMF) or a Monte Carlo set. Similarly, the "family" can be the top 10 and the bottom 10 sequences, the top 100 sequence, etc. This may also be done using error-prone PCR.

Thus, in a preferred embodiment, in silico shuffling is done using the computational methods described therein. That is, starting with either two libraries or two sequences, random recombinations of the sequences can be generated and evaluated.

In a preferred embodiment, error-prone PCR is done to generate the secondary library. See U.S. Patent Nos. 5,605,793, 5,811,238, and 5,830,721, all of which are hereby incorporated by reference. This can be done on the optimal sequence or on top members of the library, or some other artificial set or family. In this embodiment, the gene for the optimal sequence found in the computational screen of the primary library can be synthesized. Error prone PCR is then performed on the optimal sequence gene in the presence of oligonucleotides that code for the mutations at the variant positions of the secondary library (bias oligonucleotides). The addition of the oligonucleotides will create a bias favoring the incorporation of the mutations in the secondary library. Alternatively, only oligonucleotides for certain mutations may be used to bias the library.

In a preferred embodiment, gene shuffling with error prone PCR can be performed on the gene for the optimal sequence, in the presence of bias oligonucleotides, to create a DNA sequence library that

reflects the proportion of the mutations found in the secondary library. The choice of the bias oligonucleotides can be done in a variety of ways; they can chosen on the basis of their frequency, i.e. oligonucleotides encoding high mutational frequency positions can be used; alternatively, oligonucleotides containing the most variable positions can be used, such that the diversity is increased; if the secondary library is ranked, some number of top scoring positions can be used to generate bias oligonucleotides; random positions may be chosen; a few top scoring and a few low scoring ones may be chosen; etc. What is important is to generate new sequences based on preferred variable positions and sequences.

In a preferred embodiment, PCR using a wild type gene or other gene can be used, as is schematically depicted in Figure 5. In this embodiment, a starting gene is used; generally, although this is not required, the gene is the wild type gene. In some cases it may be the gene encoding the global optimized sequence, or any other sequence of the list. In this embodiment, oligonucleotides are used that correspond to the variant positions and contain the different amino acids of the secondary library. PCR is done using PCR primers at the termini, as is known in the art. This provides two benefits; the first is that this generally requires fewer oligonucleotides and can result in fewer errors. In addition, it has experimental advantages in that if the wild type gene is used, it need not be synthesized.

In a preferred embodiment, a variety of additional steps may be done to one or more secondary libraries; for example, further computational processing can occur, secondary libraries can be recombined, or cutoffs from different secondary libraries can be combined. In a preferred embodiment, a secondary library may be computationally remanipulated to form an additional secondary library (sometimes referred to herein as "tertiary libraries"). For example, any of the secondary library sequences may be chosen for a second round of PDA, by freezing or fixing some or all of the changed positions in the first secondary library. Alternatively, only changes seen in the last probability distribution table are allowed. Alternatively, the stringency of the probability table may be altered, either by increasing or decreasing the cutoff for inclusion. Similarly, the secondary library may be recombined experimentally after the first round; for example, the best gene/genes from the first screen may be taken and gene assembly redone (using techniques outlined below, multiple PCR, error prone PCR, shuffling, etc.). Alternatively, the fragments from one or more good gene(s) to change probabilities at some positions. This biases the search to an area of sequence space found in the first round of computational and experimental screening.

In a preferred embodiment, a tertiary libarary can be generated from combining secondary libraries. For example, a probability distribution table from a secondary library can be generated and recombined, wither computationally or experimentally, as outlined herein. A PDA secondary library

may be combined with a sequence alignment secondary library, and either recombined (again, computationally or experimentally) or just the cutoffs from each joined to make a new tertiary library. The top sequences from several libraries can be recombined. Primary and secondary libraries can similarly be combined. Sequences from the top of a library can be combined with sequences from the bottom of the library to more broadly sample sequence space, or only sequences distant from the top of the library can be combined. Primary and/or secondary libraries that analyzed different parts of a protein can be combined to a tertiary library that treats the combined parts of the protein. These combinations can be done to analyze large proteins, especially large multidomain proteins or complete protesomes.

In a preferred embodiment, a tertiary library can be generated using correlations in the secondary library. That is, a residue at a first variable position may be correlated to a residue at second variable position (or correlated to residues at additional positions as well). For example, two variable positions may sterically or electrostatically interact, such that if the first residue is X, the second residue must be Y. This may be either a positive or negative correlation. This correlation, or "cluster" of residues, may be both detected and used in a variety of ways. (For the genration of correlations, see the earlier cited art).

In addition, primary and secondary libraries can be combined to form new libaries; these can be random combinations or the libraries, combining the "top" sequences, or weighting the combinations (positions or residues from the first library are scored higher than those of the second library).

As outlined herein, any number of protein attributes may be altered in these methods, including, but not limited to, enzyme activity, stability, solubility, aggregation, binding affinity, binding specificity, substrate specificity, structural integrity, immunogenicity, toxicity, generate peptide and peptidomimmetic libraries, create new antibody CDR's, generate new DNA, RNA bindings, etc.

It should be noted that therapeutic proteins utilized in these methods will preferentially have residues in the hydrophobic cores screened, to prevent changes in the molecular surface of the protein that might induce immunogenic responses. Therapeutic proteins can also be designed in the region surrounding their binding sites to their receptors. Such a region can be defined, for example, by including in the design all residues within a certain distance, for example 4.5 Å of the binding site residues. This range can vary from 4 to 6-10 Å. This design will serve to improve activity and specificity.

In addition, a step method can be done; see Zhao et al., Nature Biotech. 16:258 (1998), hereby incorporated by reference.

In a preferred embodiment, the methods of the invention are used not on known scaffold proteins, but on random peptides, to search a virtual library for those sequences likely to adapt a stable conformation. As discussed above, there is a current benefit and focus on screening random peptide libraries to find novel binding/modulators. However, the sequences in these experimental libraries can be randomized at specific sites only, or throughout the sequence. The number of sequences that can be searched in these libraries grows expontentially with the number of positions that are randomized. Generally, only up to 10^{12} - 10^{15} sequences can be contained in a library because of the physical constraints of laboratories (the size of the instruments, the cost of producing large numbers of biopolymers, etc.). Other practical considerations can often limit the size of the libraries to 10^6 or fewer. These limits are reached for only 10 amino acid positions. Therefore, only a sparse sampling of sequences is possible in the search for improved proteins or peptides in experimental sequence libraries, lowering the chance of success and almost certainly missing desirable candidates. Because of the randomness of the changes in these sequences, most of the candidates in the library are not suitable, resulting in a waste of most of the effort in producing the library.

However, using the automated protein design techniques outlined herein, virtual libraries of protein sequences can be generated that are vastly larger than experimental libraries. Up to 10^{75} candidate sequences (or more) can be screened computationally and those that meet design criteria which favor stable and functional proteins can be readily selected. An experimental library consisting of the favorable candidates found in the virtual library screening can then be generated, resulting in a much more efficient use of the experimental library and overcoming the limitations of random protein libraries. Thus, the methods of the invention allow the virtual screening of a set of random peptides for peptides likely to take on a particular structure, and thus eliminating the large number of unpreferred or unallowed conformations without having to make and test the peptides.

In addition, it is possible to randomize regions or domains of protein as well.

Thus, in a preferred embodiment, the invention provides libraries of completely defined set of variant scaffold proteins, wherein at least 85% of the possible members are in the library, with at least about 90% and 95% being particularly preferred. However, it is also possible that errors are introduced into the libraries experimentally, and thus the libraries contain preferably less than 25% non-defined (e.g. error) sequences; with less than 10%, less than 5% and less than 1% particularly preferred. Thus libraries that have all members as well as some error members, or some members as well as error members are included herein.

As mentioned above, two principle benefits come from the virtual library screening: (1) the automated protein design generates a list of sequence candidates that are favored to meet design criteria; it also shows which positions in the sequence are readily changed and which positions are unlikely to change without disrupting protein stability and function. An experimental random library can be generated that

is only randomized at the readily changeable, non-disruptive sequence positions. (2) The diversity of amino acids at these positions can be limited to those that the automated design shows are compatible with these positions. Thus, by limiting the number of randomized positions and the number of possibilities at these positions, the number of wasted sequences produced in the experimental library is reduced, thereby increasing the probability of success in finding sequences with useful properties.

For example, the table below lists the 10 favored sequences candidates from the virtual screening of 12 positions in a protein. It shows that positions 9, 10 and 12 are most likely to have changes that do not disrupt the function of the protein, suggesting that a random experimental library that randomizes positions 9, 10 and 12 will have a higher fraction of desirable sequences. Also, the virtual library suggests that position 10 is most compatible with Ile or Phe residues, further limiting the size of the library and allowing a more complete screening of good sequences.

	1	2	3	4	5	6	7	8	9	10	11	12
1	LEU	LEU	ILE	ILE	ALA	LEU	LEU	LEU	LEU	PHE	ALA	LEU
2	LEU	LEU	ILE	ILE	ALA	LEU	LEU	LEU	LEU	ILE	ALA	LEU
3	LEU	LEU	ILE	ILE	ALA	LEU	LEU	LEU	LEU	ILE	ALA	LEU
4	LEU	LEU	ILE	ILE	ALA	LEU	LEU	LEU	LEU	PHE	ALA	ILE
5	LEU	LEU	ILE	ILE	ALA	LEU	LEU	LEU	LEU	PHE	ALA	ILE
6	LEU	LEU	ILE	ILE	ALA	LEU	LEU	LEU	LEU	ILE	ALA	ILE
7	LEU	LEU	ILE	ILE	ALA	LEU	LEU	LEU	ILE	PHE	ALA	LEU
8	LEU	LEU	ILE	ILE	ALA	LEU	LEU	LEU	LEU	ILE	ALA	ILE
9	LEU	LEU	ILE	ILE	ALA	LEU	LEU	LEU	ILE	PHE	ALA	LEU
10	LEU	LEU	ILE	ILE	ALA	LEU	LEU	LEU	LEU	LEU	ALA	LEU

The automated design method uses physical chemical criteria to screen sequences, resulting in sequences that are likely to be stable, structured, and that preserve function, if needed. Different design criteria can be used to produce candidate sets that are biased for properties such as charged, solubility, or active site characteristics (polarity, size), or are biased to have certain amino acids at certain positions. That is, The candidate bioactive agents and candidate nucleic acids are randomized, either fully randomized or they are biased in their randomization, e.g. in nucleotide/residue frequency generally or per position. By "randomized" or grammatical equivalents herein is meant that each nucleic acid and peptide consists of essentially random nucleotides and amino acids, respectively. Thus, any amino acid residue may be incorporated at any position. The synthetic process can be designed to generate randomized peptides and/or nucleic acids, to allow the formation of all or most of the possible combinations over the length of the nucleic acid, thus forming a library of randomized candidate nucleic acids.

In one embodiment, the library is fully randomized, with no sequence preferences or constants at any position. In a preferred embodiment, the library is biased. That is, some positions within the sequence are either held constant, or are selected from a limited number of possibilities. For example, in a preferred embodiment, the nucleotides or amino acid residues are randomized within a defined class, for example, of hydrophobic amino acids, hydrophilic residues, sterically biased (either small or large) residues, towards the creation of cysteines, for cross-linking, prolines for SH-3 domains, serines, threonines, tyrosines or histidines for phosphorylation sites, etc., or to purines, etc.

In a preferred embodiment, the bias is towards peptides or nucleic acids that interact with known classes of molecules. For example, it is known that much of intracellular signaling is carried out via short regions of polypeptides interacting with other polypeptides through small peptide domains. For instance, a short region from the HIV-1 envelope cytoplasmic domain has been previously shown to block the action of cellular calmodulin. Regions of the Fas cytoplasmic domain, which shows homology to the mastoparan toxin from Wasps, can be limited to a short peptide region with death-inducing apoptotic or G protein inducing functions. Magainin, a natural peptide derived from Xenopus, can have potent anti-tumour and anti-microbial activity. Short peptide fragments of a protein kinase C isozyme (ßPKC), have been shown to block nuclear translocation of ßPKC in Xenopus oocytes following stimulation. And, short SH-3 target peptides have been used as psuedosubstrates for specific binding to SH-3 proteins. This is of course a short list of available peptides with biological activity, as the literature is dense in this area. Thus, there is much precedent for the potential of small peptides to have activity on intracellular signaling cascades. In addition, agonists and antagonists of any number of molecules may be used as the basis of biased randomization of candidate bioactive agents as well.

In general, the generation of a prescreened random peptide libraries may be described as follows. Any structure, whether a known structure, for example a portion of a known protein, a known peptide, etc., or a synthetic structure, can be used as the backbone for PDA. For example, structures from X-ray crystallographic techniques, NMR techniques, de novo modelling, homology modelling, etc. may all be used to pick a backbone for which sequences are desired. Similarly, a number of molecules or protein domains are suitable as starting points for the generation of biased randomized candidate bioactive agents. A large number of small molecule domains are known, that confer a common function, structure or affinity. In addition, as is appreciated in the art, areas of weak amino acid homology may have strong structural homology. A number of these molecules, domains, and/or corresponding consensus sequences, are known, including, but are not limited to, SH-2 domains, SH-3 domains, Pleckstrin, death domains, protease cleavage/recognition sites, enzyme inhibitors, enzyme substrates, Traf, etc. Similarly, there are a number of known nucleic acid binding proteins containing domains suitable for use in the invention. For example, leucine zipper consensus sequences are known.

Thus, in general, known peptide ligands can be used as the starting backbone for the generation of the primary library.

In addition, structures known to take on certain conformations may be used to create a backbone, and then sequences screened for those that are likely to take on that conformation. For example, there are a wide variety of "ministructures" known, sometimes referred to as "presentation structures", that can confer conformational stability or give a random sequence a conformationally restricted form. Proteins interact with each other largely through conformationally constrained domains. Although small peptides with freely rotating amino and carboxyl termini can have potent functions as is known in the art, the conversion of such peptide structures into pharmacologic agents is difficult due to the inability to predict side-chain positions for peptidomimetic synthesis. Therefore the presentation of peptides in conformationally constrained structures will benefit both the later generation of pharmaceuticals and will also likely lead to higher affinity interactions of the peptide with the target protein. This fact has been recognized in the combinatorial library generation systems using biologically generated short peptides in bacterial phage systems. A number of workers have constructed small domain molecules in which one might present randomized peptide structures.

Thus, synthetic presentation structures, i.e. artificial polypeptides, are capable of presenting a randomized peptide as a conformationally-restricted domain. Preferred presentation structures maximize accessibility to the peptide by presenting it on an exterior loop. Accordingly, suitable presentation structures include, but are not limited to, minibody structures, loops on beta-sheet turns and coiled-coil stem structures in which residues not critical to structure are randomized, zinc-finger domains, cysteine-linked (disulfide) structures, transglutaminase linked structures, cyclic peptides, B-loop structures, helical barrels or bundles, leucine zipper motifs, etc.

In a preferred embodiment, the presentation structure is a coiled-coil structure, allowing the presentation of the randomized peptide on an exterior loop. See, for example, Myszka et al., Biochem. 33:2362-2373 (1994), hereby incorporated by reference, and Figure 3). Using this system investigators have isolated peptides capable of high affinity interaction with the appropriate target. In general, coiled-coil structures allow for between 6 to 20 randomized positions; (see Martin et al., EMBO J. 13(22):5303-5309 (1994), incorporated by reference).

In a preferred embodiment, the presentation structure is a minibody structure. A "minibody" is essentially composed of a minimal antibody complementarity region. The minibody presentation structure generally provides two randomizing regions that in the folded protein are presented along a single face of the tertiary structure. See for example Bianchi et al., J. Mol. Biol. 236(2):649-59 (1994), and references cited therein, all of which are incorporated by reference). Investigators have shown this minimal domain is stable in solution and have used phage selection systems in combinatorial